



Grid y Computación  
de Altas Prestaciones

**GRyCAP**

# RETOS COMPUTACIONALES EN LA MEDICINA PERSONALIZADA

**Ignacio Blanquer**  
**Vicente Hernández**



UNIVERSITAT  
POLITÈCNICA  
DE VALÈNCIA



Instituto de Instrumentación  
para Imagen Molecular



- Comunidad Acostumbrada a Trabajar de Forma Colaborativa en Internet
  - GeneBank, PDB, SWISSPROT, KEGG, Portal de NCBI.
- Usuarios de Muchas Herramientas Comunes
  - BLAST, ClustalW, ePCR, SAGE, EMBOSS , Phylip, ...
- Una Comunidad Abierta y Partidaria del Código Abierto y el Acceso Abierto a los Datos.
- Consumidores de ~10% de los Recursos Computacionales de las Infraestructuras Europeas de Investigación.



# ESTA PERCEPCIÓN NO ES SÓLO DESDE LAS TIC



**GRyCAP**  
Grid y Computación de Altas Prestaciones

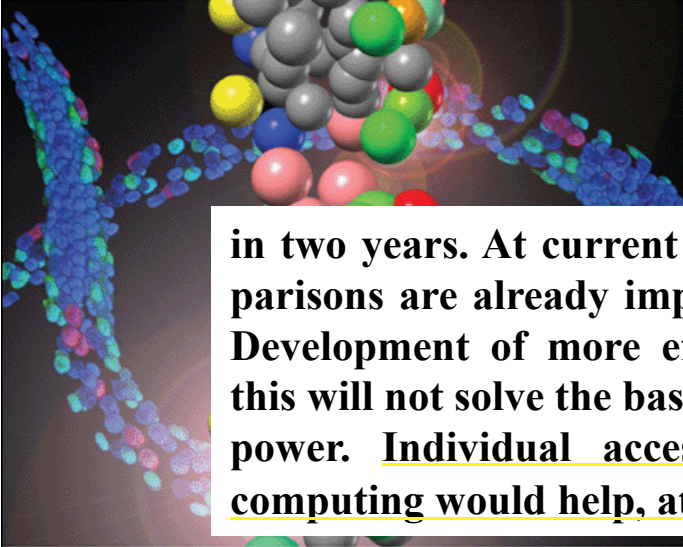
www.grycap.upv.es

EDITORIAL

September 2009 | volume 6 | number 9

**nature | methods**

www.nature.com/naturemethods Techniques for life scientists and chemists



**in two years. At current database sizes all-versus-all comparisons are already impossible without a supercomputer. Development of more efficient algorithms will help, but this will not solve the basic problem of too little computing power. Individual access to supercomputers or cloud computing would help, at least temporarily.**

- A digital atlas of the worm
- Tools for metagenomics
- A reporter line for erythroid differentiation
- BreakDancer
- Addressing the crystallography phase problem

## Metagenomics versus Moore's law

Metagenomics sprang from advances in sequencing technology, and continued improvements are providing data in quantities unimaginable a few years ago. But without concerted efforts, the amount of data will quickly outpace the ability of scientists to analyze it.

As Craig Venter sails the oceans collecting seawater samples to profile microbial communities by high-throughput sequence analysis, microbiologists around the world are busy collecting their own samples. The diversity of locations—from Antarctic lakes to human armpits—highlights the reality that microscopic organisms represent a significant fraction of the Earth's ecosystem.

Any population in this large is certain to have profound influences on its environment. Yet our knowledge of

40 megabases. Today there are over 4,000 sequenced metagenomes, and their size and number are increasing. Each new pyrosequenced metagenome is 200–500 megabases, and those generated on Illumina platforms are 20–50 gigabases. To analyze these metagenomes using established pipelines would take tens of years on a single processor and weeks to months on machines with up to 1,000 processors. The rate of increase in sequence generation is far outpacing Moore's law, and the cost of analyzing the largest datasets

ally gene  
erations,  
enomes  
ting this  
n genes,  
on genes  
all com-  
puter.  
isly but  
omput-  
or cloud

o avoid  
available  
subtyp-  
ic anno-  
it does  
to hap-  
biology  
set sizes  
sed sup-  
needs to

SOLID platforms. Most metagenomics analysis pipelines are designed for Sanger sequencing data, so the short read lengths and error profiles of the new methods present challenges for data analysis and interpretation.

Reports on pages 639 and 673 and an accompanying News and Views on page 636 illustrate some of the dangers and challenges involved and describe new algorithms to deal with them. More work is needed to assess the new technologies and develop optimized analysis pipelines, and these efforts are well underway.

But even as these problems are being solved, a larger problem has taken the community off-guard: the exponentially increasing amount of sequence data. Just over three years ago, the first two second-generation sequencing platform-based shotgun metagenomes were reported—each less than

decrease computational demands by improving data sharing through standards and centralized coordination and by aggregating computationally intensive operations.

This summer, after discussions at the International Conference on Systems for Intelligent Molecular Biology, community members formed the M5 (metagenomics, metadata, metaanalysis, multiscale-models and metainfrastructure) Consortium under the roof of the Genomics Standards Consortium to devise a solution to the coming gridlock. Their proposed 'M5 Platform'—to be announced later this year—deserves the support of the community, funding agencies and those who hold the keys to the high-performance computing centers. Unless major efforts are taken immediately, researchers will find they have a wealth of data but no way to interpret it.

# INFRAESTRUCTURAS COMPUTACIONALES



**GRyCAP**

Grid y Computación de Altas Prestaciones

[www.grycap.upv.es](http://www.grycap.upv.es)



**Google  
App Engine**



 **Windows Azure™**



# EXPERIMENTOS A GRAN ESCALA

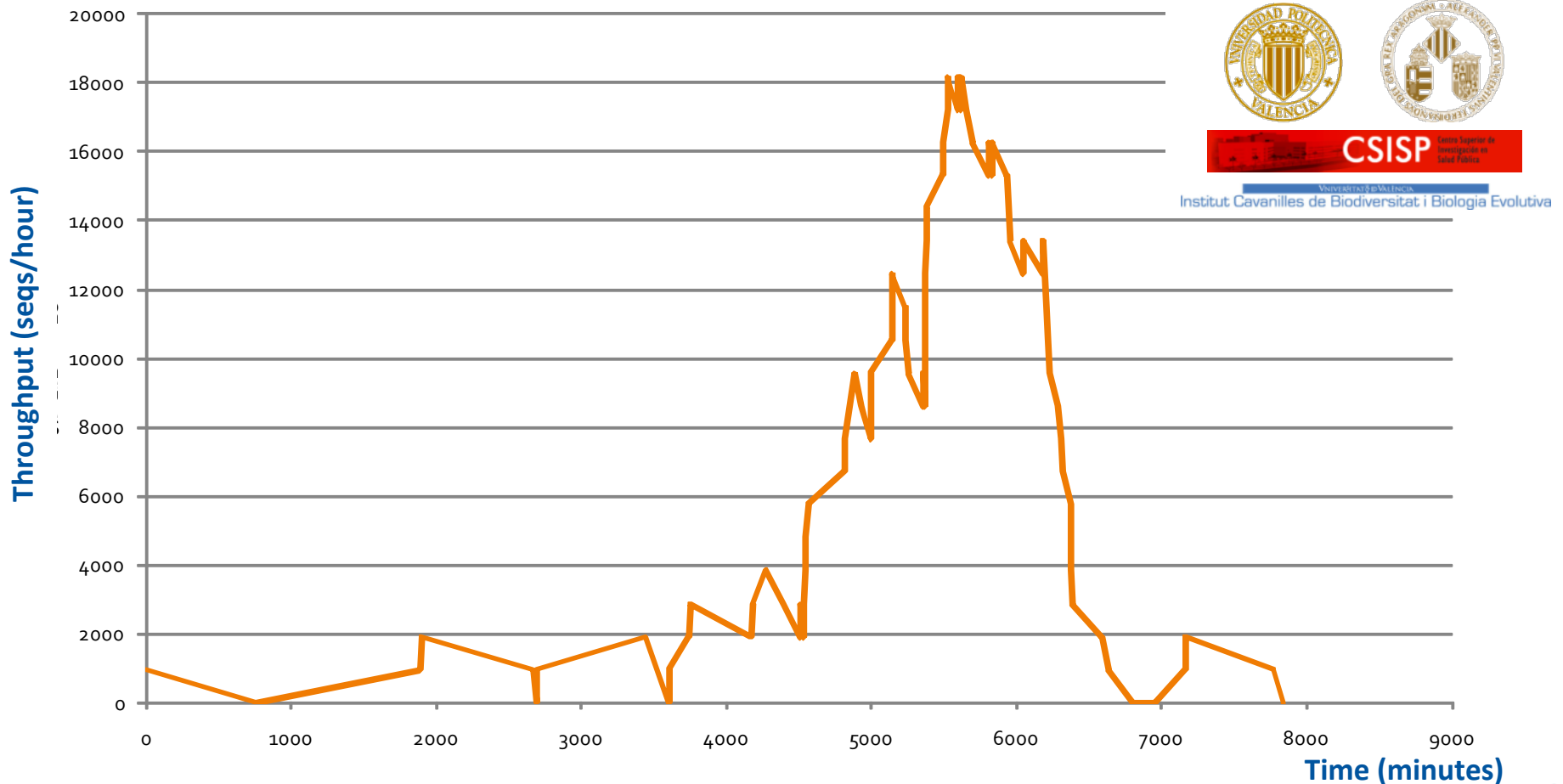
## Bioinformatics Use Case



**GRyCAP**  
Grid y Computación de Altas Prestaciones

[www.grycap.upv.es](http://www.grycap.upv.es)

- 600 Computadores Usados Simultáneamente en EGI.







- A Pesar del Crecimiento que las Bases de Datos de Información Genética, Genómica y Proteómica no hay Cambios en la Estructura Original.
- Por ello, Tanto la Eficiencia Como la Coherencia de los Contenidos se Debilitan a Medida que su Tamaño Aumenta Exponencialmente.
- Una Aplicación Bioinformática Moderna Exige un Acoplamiento Fino entre Diseño Correcto de Datos y Mecanismos Potentes de Búsqueda.
  - Estos Contenidos Deben de Explotarse con Herramientas de Búsqueda con la Potencia Computacional Necesaria.



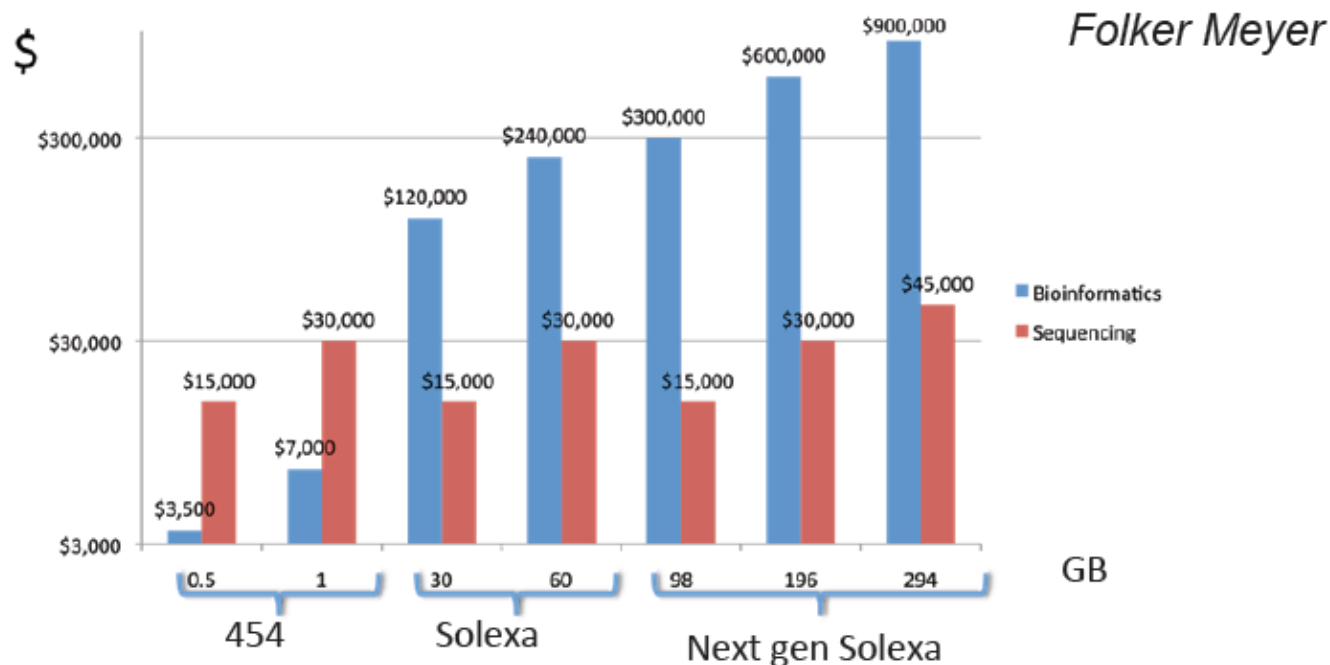
# PERO LAS COSAS SE COMPLICAN



**GRyCAP**  
Grid y Computación de Altas Prestaciones

[www.grycap.upv.es](http://www.grycap.upv.es)

the “simple” problem: sequencing outpaces moore’s law



- Values are for BLAST searches on Amazon EC2 (from *Wilkening et al, IEEE Cluster09*)
- Acknowledged: BLAST is not the ideal tool for a lot of this, but...

The image shows two overlapping web browser windows. The background window is Internet Explorer displaying a blog post from 'Pathogens: Genes and Genomes' by the Beijing Genomics Institute. The foreground window is also Internet Explorer, showing the 'Genome 10K Project' website. The website features a navigation bar with links to Database, News, Events, Participants, Publications, and For G10KCOS. A large banner with a DNA double helix and the text 'GENOME 10K Unveiling animal diversity' is prominent. A sidebar on the right lists options like 'Become a participant', 'Supporters', and 'Postdoc Position'. The main content area is titled 'Genome 10K Project' and describes the project's goal of assembling a genomic zoo of 10,000 vertebrate species. An 'Accomplishments' box on the right highlights the project's progress.

**Pathogens: Genes and Genomes**  
A heady mix of bacterial pathogenomics, next-generation sequencing, and more.

**Archive**  
Posts Tagged 'beijing genomics institute'

**New Illumina Announced – BGI to become**  
January 13th, 2010 Nick Loman

The ritual of checking Twitter with my coffee this morning was interrupted by a [chattering](#) about the new Illumina sequencer, the HiSeq. I won't cover this in detail as there are already [Daniel McArthur](#), [David Dooling](#) and [Genomics Lawyer](#) machine, \$10,000 for the reagents, 2 x 100 base-pair sequence data per run. And a terrible name.

Technology wise this has been done through parallelism. Illumina have increased the number of flow-cells in each lane, and the readable area of the flow-cell has been increased. The market where Intel have managed to keep pace with processor cores on each chip.

The other big news is that [Beijing Genomics Institute](#) will hand them the title of largest genome center. To the [map of high-throughput sequencers](#), the Broad Institute doubt we will see a bunch of other genome centres pl...

**Update:** For those with GA2s looking to upgrade, I can give you a paltry \$150,000 off the list price of the HiSeq.

[High-throughput sequencing](#) [beijing genomics](#)

**Genome 10K Project**

The Genome 10K project aims to assemble a genomic zoo—a collection of DNA sequences representing the genomes of 10,000 vertebrate species, approximately one for every vertebrate genus. The trajectory of cost reduction in DNA sequencing suggests that this project will be feasible within a few years. Capturing the genetic diversity of vertebrate species would create an unprecedented resource for the life sciences and for worldwide conservation efforts.

The growing Genome 10K Community of Scientists (G10KCOS), made up of leading scientists representing major zoos, museums, research centers, and universities around the world, is dedicated to coordinating efforts in tissue specimen collection that will lay the groundwork for a large-scale sequencing and analysis project.

**Accomplishments**

- The Genome 10K database catalogs specimens from more than 16,000 vertebrate species, living and recently extinct. The species include mammals, birds, nonavian reptiles, amphibians, and fishes, many of which are threatened or endangered.
- Inaugural publication in the



# CONCLUSIONES

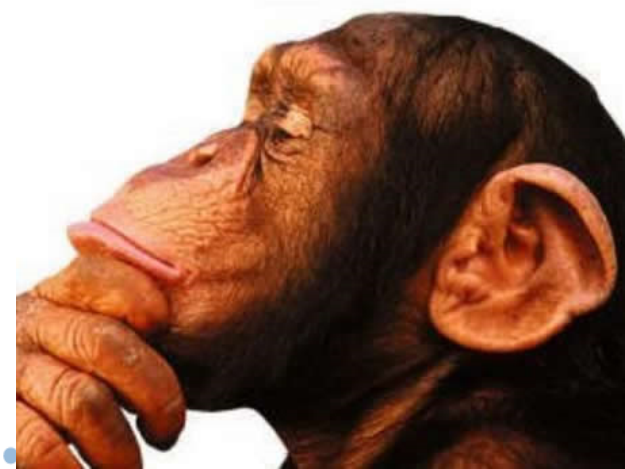


**GRyCAP**

Grid y Computación de Altas Prestaciones

[www.grycap.upv.es](http://www.grycap.upv.es)

- La Genómica puede Convertirse en la Actividad con Mayor Demanda Computacional y de Almacenamiento en la Industria y la Investigación.
- Las Técnicas de Cloud Computing se ven Como una Solución, al Menos a Corto Plazo.
- Las Infraestructuras Públicas de Investigación y los Proveedores Comerciales Jugarán un Papel Fundamental en la Medicina Personalizada.



Ignacio Blanquer  
Vicente Hernández  
Universidad Politécnica de Valencia  
Camino de Vera s/n,  
46022 Valencia  
Tel: +34-963879743  
Fax. +34-963877274  
E-mail: [iblanque@dsic.upv.es](mailto:iblanque@dsic.upv.es)  
[vhernand@dsic.upv.es](mailto:vhernand@dsic.upv.es)

